Our favourite alternative splice site.

Hubert Lerivray, Agnès Méreau, Howard Beverley Osborne

Our Favourite Alternative Splice site

Hubert Lerivray, Agnès Méreau and H. Beverley Osborne

Groupe Régulation de l'Expression Génétique, UMR 6061 Génétique et Développement, CNRS/Université de Rennes I, IFR 140 Génétique Fonctionnelle, Agronomie et Santé, Faculté de Médecine, 2 avenue du Pr. Léon Bernard, CS 34317, Rennes cedex, 35043, France

Corresponding author: H.Lerivray, UMR 6061 Génétique et Développement, CNRS, Université de Rennes I, Faculté de Médecine, 2 avenue du Pr. Léon Bernard, CS 34317, Rennes cedex, 35043, France.

email hubert.lerivray@univ-rennes1.fr; Tel: 33-(0)2-23-23-44-64, Fax: 33-(0)2-23-23-44-78

Running title: Alternative Splice web sites

Key words: database, alternative splicing, transcription, genomics

Abbreviations used: Alternative Splicing Annotation Project (ASAP); Alternative Splicing Database (ASD); Extended Alternatively Spliced EST Database (EASED); Expressed Sequence Tag (EST); Friendly Alternative Splicing and Transcripts DataBase (FAST DB); Manually Annotated Alternative Spliced Events Database (MAASE); Untranslated Region (UTR).

Abstract

Alternative splicing is a widespread mechanism in mammals that generates several mRNAs from one gene, thereby creating genetic diversity of the genome. Variant splice patterns are often specific to different stages of development or particular tissues and alternative splicing defects are more and more frequently detected in genetic diseases or cancers. This increasingly important position occupied by alternative splicing in the function and the regulation of cellular process, makes it critical to have an easy to use data repository for the biological and medical research communities. We have compared web resources that give access to information on alternative spliced genes, the “Friendly Alternative Splicing and Transcripts DataBase” (FAST DB) site came out as our favourite.

Introduction

Alternative splicing of mRNA allows several products and consequently several protein isoforms with different functions, to be synthesized from a unique gene. It is a major mechanism for increasing transcriptome diversity in metazoan genomes. Initially considered as a marginal process affecting less than 5% of all genes (Sharp, 1994), it became clear, through several bioinformatic or DNA microarray studies, that it was highly abundant and concerned 50-70% of human genes (Modrek and Lee, 2002; Johnson et al., 2003) producing many isoforms with different functions. Moreover, it is now clear that many human diseases are caused by alternative splicing deficiencies (Krawczak et al., 2000; Musunuru, 2003; Ranum and Day, 2004) and alternative splicing modifications are now major targets in medical therapy (Sazani and Kole, 2003; Garcia-Blanco et al., 2004).

Cell biologists studying their favourite gene product may become confronted with the question of protein isoforms produced by alternative splicing of a single pre-mRNA or from pre-mRNAs transcribed by different promoters within the same gene. Access via the Internet to databases with user-friendly tools then becomes a premium necessity. Furthermore, recent data on the relationship between splicing factors, promoter structure and transcription elongation (Cramer et al, 1997; Fong and Zhou, 2001; Furger et al., 2002; Auboeuf et al., 2004) support the concept that splicing and transcription are functionally coupled and that this coupling is an additional level of regulation of alternative splicing (Kornblihtt et al., 2004). The capacity of an interface to integrate this aspect should also be taken into account when choosing a web resource to use.

Some available Alternative Splicing site web based databases.

A large number of databases on alternative splicing have been published so far and, to the non-initiated, the task of finding one to fit his or her needs can be daunting. We took a simple approach to making a repertoire of available sites, a Google search using the keywords "Alternative Splicing database". This produced more than 10 pages of web sites. From this initial search 8 sites were retained for further consideration, these are listed in Table 1. The choice was made on the basis of availability (some listed sites are no longer active) and a minimum of data analysis tools.

Sequence data.

In Table 1 are briefly summarised, for each site, the source of sequence data and the organisms concerned. The sequence data from a number of different sources were used to generate the listed databases. All are based on the fundamental principal of computationally determined splice events deduced through the alignments of mRNA derived sequences with genomic sequences. The mRNA derived sequences are of two origins, experimental published cDNAs sequences or EST (Expressed Sequence Tag) data. Several of the available databases such as European Bioinformatics Institute Alternative Splicing Database (ASD) (Thanaraj et al., 2004), the University of California San Diego Manually Annotated Alternative Spliced Events Database (MAASE) and the recently published Friendly Alternative Splicing and Transcripts DataBase (FAST DB) (de la Grange et al., 2005) use cDNA sequences from public databanks. The ASD site in fact interacts with three databases one of which (AEdb) contains curated information. The database for the MAASE site contains only a limited number of human and mouse genes. The FAST DB site uses primarily the cDNA data but access to the EST data is proposed. Most of the other sites use EST data sources.

Typically ESTs are produced in large batches from a given cDNA library and constitute tags (some coding, others not) of expressed genes. As tags, not all splice events are presents in EST data and, in addition, they may contain sequence errors and cDNAs corresponding to non-specific (background) splicing events. Caution is therefore required when the EST data has not been manually verified. However, the cDNA libraries used to obtain the EST data are made from different tissue. Hence they can give primary information on gene and splice variant expression patterns. The Prosplice site is worth a special mention as it appears to be the only one which also uses protein sequence data.

Updating of data is also an important aspect. Although we did not catalogue this information we did note that several of the sites listed in Table 1 had not been updated since 2002. In databases using published experimental data or in curated nucleotide databases, the number of genes is still limited as the splicing patterns of many genes are not yet well

characterised. Accordingly, even for currently updated databases, it might not be possible to find one's favourite gene.

Finding your gene and what you get.

One of the important aspects for a user when entering a web site is to get started easily. For all the sites the search terms Accession number or Acronym will find the desired gene product. Therefore, as a criteria of facility we took the ability of a site to find our favourite gene (alpha tropomyosin) just by typing in "tropomyosin". The ASD, Alternative Splicing Annotation Project (ASAP) (Lee et al. 2003), Extended Alternatively Spliced EST Database (EASED) (Pospisel et al; 2004) and FAST DB site all responded favourably. Notably, "alpha tropomyosin" was not recognised by the ASAP, EASED and FAST DB sites whereas the ASD site interpreted the combined term as "alpha" OR "tropomyosin".

Many of the databases appear to have been conceived to specifically address certain aspects of gene structure. Most contain links to several application programs. Depending on the website interface, it is possible to generate the alternatively spliced variants, intron-exon structure and often to analyse them in terms of expression. For some sites these aspects are more convivial than others to the non-genomic orientated user. In figure 1, examples of alpha tropomyosin (TPM1) outputs in EASED, ASAP and FAST DB websites are presented. The EASED site returns a first reply page with a list of accession numbers, clicking on one gives a second page (Fig 1A) with diagrams of splice pattern and much more data including tissue and developmental stage expression. However, these are colour coded and interpretation by colour blind people is difficult. For ASAP the entry appears first as an overview of the spliced gene which is link to a page with the various mRNA isoforms of the considered gene (Fig1B). The ASD site, which interrogates three European Bioinformatics Institute databases, also returns a list of possible gene products from a wide range of species (not shown). With FAST DB, validating "Tropomyosin 1 (alpha Tropomyosin)" in the search results sends the user to the main page that has much synthetic information starting by a map of the gene showing all the known exons and introns (Fig 1C). Furthermore, it is possible to get exon or intron sequences, in another frame of the same window, by clicking on the wanted gene region. As far as we could determine this was the only site that is directly linked to web application programmes that can predict promoters, or analyse transcription termination.

Information on tissue expression patterns.

Data on the tissue expression pattern can be obtained in several ways, either experimentally by *in situ* hybridisation or by the direct analysis of mRNAs extracted from specific tissue. Information on tissue expression is also inherent in databases using EST data

obtained from cDNA libraries generated from individual tissues, but not all present it in a readily available or easily interpretable form. We already mentioned that EASED site uses a colour code that has certain drawbacks. For the ASD site one must first scroll down several pages to the diagrams showing the splice variants, then putting the mouse indicator over the splice variant indicator (SP number) brings up a window with the known expression data. For the FAST DB site it was necessary to first choose the "Transcripts view" window. Here the various known splice variants are listed diagrammatically. Against each is directly seen the Accession number and the tissue source for the mRNA used to produce the sequence data. Incidentally, there are also links to the corresponding Genbank entry and to PubMed for the publication if one exists. For most of the other sites, information on the tissue expression pattern was either absent or difficult to use when examined from a gene candidate approach.

EST and cDNA sequence data have not been obtained for every gene from every tissue. Therefore, relative to a specific gene product it is highly probable that the tissue expression pattern data will be incomplete. One way for the researcher to complete this data is by a PCR analysis of extracted mRNA using primers that will identify the different splice variants of a pre-mRNA. To our knowledge the FAST DB site is the only one which offers a tool that helps the researcher in this task; this is the *in silico* PCR module. Clicking on "*in silico* PCR" engenders the nucleotide sequence alignment of all the cDNA sequences in the main window thereby visualizing the alternative exon junctions and also the possible polymorphism. Forward and reverse primers can be chosen (copy and paste in a second window) and used for the *in silico* PCR. The result will give the size and the sequence of the products predicted to be amplified from the different transcripts, the accession numbers are also indicated. Another useful tool developed in FAST DB to analyse the expression pattern of gene products is the "probe alignment" facility. With this tool it is possible to determine what part of a known probe can match with the mRNA or genomic DNA.

Outward links.

In many instances a researcher needs to switch between Web sites dedicated to specific needs. In these instances, links programmed into the web site are an important bonus and time saver. Many of the Alternative Splice web sites have links out to the original Genbank or EMBL entries, the Unigene database, the EnsEMBL database or the Uniprot/SwissProt database. The FAST DB site has several important additional links. Researchers studying transcription use computer tools in order to predict promoter regions, transcription factor binding sites or to analyse the untranslated regions (UTR). From FAST DB it is possible to directly analyze up to 5000 nucleotides upstream of a chosen position

(intron or exon sequence) with links to these external tools. Transcription initiation can be analyzed using links to different web tools allowing promoter prediction (6 links), transcription factor binding site predictions (9 links) or 5'UTR analysis (7 links). In the same way, analysis of terminal exons and transcription termination can be done using external web tools (7 links). Furthermore, this site is not exclusive, links are provided to many other alternative splicing websites. Finally all the information produced from this site can be summarized and saved in a pdf file.

Our best choice.

To conclude we found the FAST DB site is a very friendly and easy to use website, fine-tuned to analysing alternative splicing for non-specialist researchers. The large number of links to sequence analysis tools and the *in silico* PCR facility rapidly allows association between alternative splicing and transcription. However, as with all websites the database does not contain information on all possible human or mouse genes. As mentioned abov,e this may be due to a lack of available information but also because compete coverage requires time. The website builders are still improving the site and it can be assumed that database updates and other innovations in alternative splicing studies will be introduced.

Acknowledgements

We wish to thank Dr D. Auboeuf and P. de la Grange for letting us know about FAST DB and their comments on this manuscript. We are grateful to Pr S.Hardy for his critical comments and suggestions. Work in the author's laboratory is supported by the CNRS, Ministère de l'Education Nationale, and the Association Française contre les Myopathies.

References

Auboeuf, D., Dowhan, D.H., Kang, Y.K., Larkin, K., Lee, J.W., Berget, S.M. and O'Malley, B.W. (2004) Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. Proc. Natl. Acad. Sci. U.S.A. **101,** 2270-2274.

Cramer, P., Pesce, C.G., Baralle, F.E. and Kornblihtt, A.R. (1997) Functional association between promoter structure and transcript alternative splicing. Proc. Natl. Acad. Sci. U. S. A. **94**, 11456-11460.

De la Grange, P., Dutertre, M., Martin, N. and Auboeuf, D. (2005) FAST DB: a website resource for the study of the expression regulation of human gene products. Nucleic Acids Res. **33,** 4276-4284.

Fong, Y.W. and Zhou, Q. (2001) Stimulatory effect of splicing factors on transcriptional elongation. Nature **414,** 929-33.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A. and Proudfoot, N.J. (2002) Promoter proximal splice sites enhance transcription. Genes Dev. **16**, 2792-2799.

Garcia-Blanco, M.A., Baraniak, A.P. and Lasda, E.L. (2004) Alternative splicing in disease and therapy. Nat. Biotechnol. **22**, 535-46

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. Science **302**, 2141-2144.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J. and Nogues, G. (2004) Multiple links between transcription and splicing. RNA **10**, 1489-1498.

Krawczak, M., Ball, E.V., Fenton, I., Stenson, P.D., Abeysinghe, S., Thomas, N. and Cooper D.N. (2000) Human gene mutation database-a biomedical information and research resource. Hum. Mutat. **15**, 45-51.

Lee, C., Atanelov, L., Modrek, B. and Xing, Y. (2003) ASAP: The Alternative Splicing Annotation Project. Nucleic Acids Res. **31**, 101-5.

Mordrek, B. and Lee, C. (2002) A genomic view of alternative splicing. Nature **30**, 13-19

Musunuru, K. (2003) Cell-specific RNA-binding proteins in human disease. Trends Cardiovasc Med. **13**, 188-95.

Pospisil, H., Herrmann, A., Bortfeldt, R. and Reich, J. (2004) EASED: Extended Alternatively Spliced EST Database. Nucleic Acids Res. **32**, D70-D74.

Ranum, L.P. and Day, J.W. (2004) Pathogenic RNA repeats: an expanding role in genetic disease. Trends Genet. **20**, 506-512.

Sazani, P. and Kole, R. (2003) Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. J. Clin. Invest. **4,** 481-486.

Sharp, P.A. (1994) Split genes and RNA splicing. Cell **77,** 805-815.

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V. and Muilu, J; (2004). ASD: the Alternative Splicing Database. Nucleic Acids Res. **32,** D64-D69.

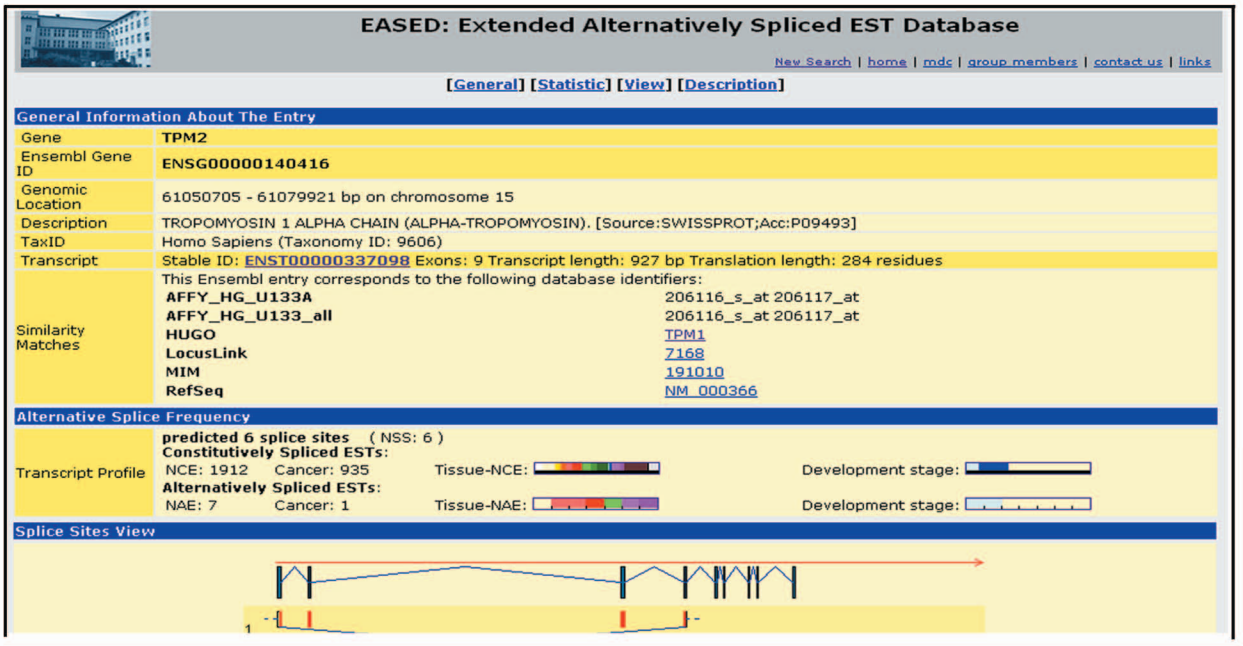
Database	Entry mode	Organism	URL
ASAP	Unigene clusters of ESTs	Human	http://bioinfo.mbi.ucla.edu/ASAP/
ASD (AltSplice-AltExtron)	Computationally created (Gene-EST/mRNA alignments)	Human and other species	http://www.ebi.ac.uk/asd//
ASD (AEdb)	Manually created from literature	All animals	http://www.ebi.ac.uk/asd//
EASED	EST and mRNA	Human, Plant, Cow, Worm, Fly, Fish, Mouse, Rat ,Frog	http://eased.bioinf.mdc-berlin.de/
FAST DB	mRNA and EST	Human and mouse	http://193.48.40.18/fastdb
SpliceNest	EST and mRNA	Human, Mouse, Drosophila, Arabidopsis	http://splicenest.molgen.mpg.de/
MAASE	Manually Annotated Alternatively Spliced Events	Human and Mouse	http://splice.sdsc.edu/
ProSplicer	Protein, mRNA, and EST	Human	http://prosplicer.mbc.nctu.edu.tw/

Table 1: **Some database resources on alternative splicing**

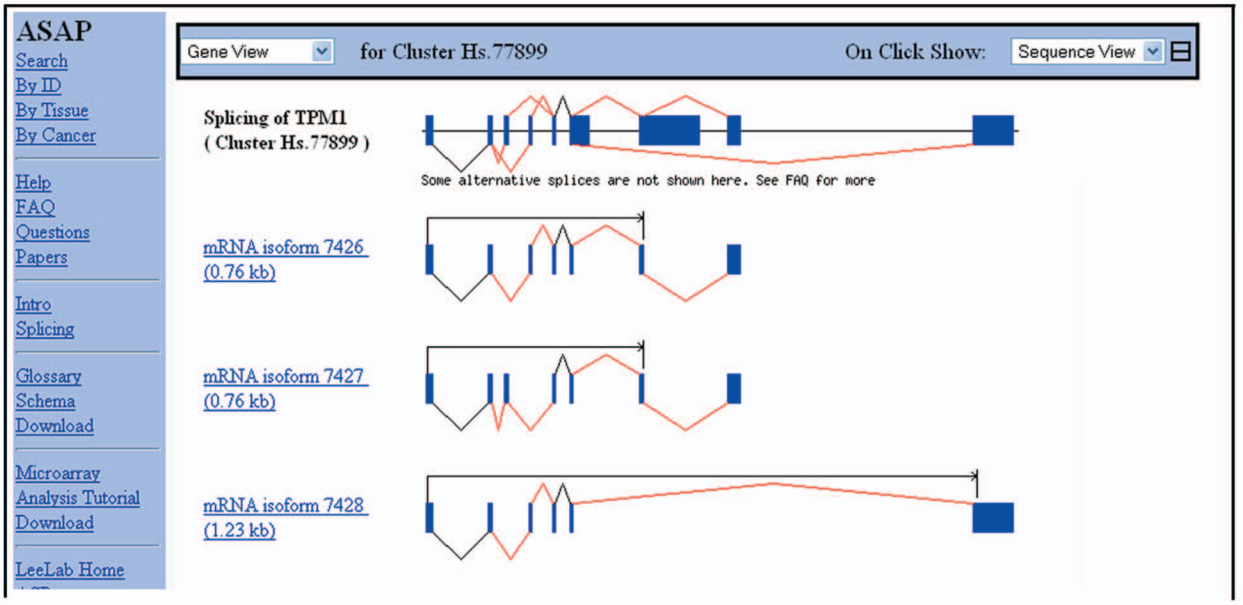
Listed in column 1 are some of the databases that were analysed. Sites that did not produce results for several genes known to undergo alternative splicing are not listed. The second and third columns (Entry mode and Organism) indicate the type and species source of the mRNA derived sequence data that was compared to the genomic sequence data.

Figure 1: Examples of outputs showing part of the pages illustrating the patterns of alternative splicing from EASED (A), ASAP (B) and FAST DB (C), websites using "TPM1" (alpha tropomyosin) as entry term.

A



B



C